December 20, 2024

E-Filed

The Honorable Thomas S. Hixson
United States District Court for the Northern District of California
San Francisco Courthouse, Courtroom E – 15th Floor
450 Golden Gate Avenue
San Francisco, CA 94102

Re: *Kadrey, et al v. Meta Platforms, Inc.*; Case No. 23-cv-03417-VC (TSH)

Dear Judge Hixson:

Plaintiffs and Defendant Meta Platforms, Inc. ("Meta") jointly submit this letter brief regarding Plaintiffs' request for an order compelling Meta's production of the documents and data described herein. The parties met and conferred on December 17, but were unable to reach a resolution.

I. PLAINTIFFS' STATEMENT

In his recent deposition, Meta research engineer Nikolay Bashlykov testified to both the existence and the location of key data underlying the Llama Models' training datasets of copyrighted works and copyright mitigation functions. Subsequent communications with Meta revealed that *none* of this data was produced to Plaintiffs or their experts. It should be. Plaintiffs thus respectfully request that the Court order Meta to produce the following three categories of data.

A. Torrenting Data

Meta's torrenting of pirated works has become a topic of much attention in this case. *See*, *e.g.*, Dkt. Nos. 300; 308; 321; 335; 342. Once Meta's witnesses began testifying about this practice, it became clear that Meta's copyright infringement was far more brazen than Plaintiffs previously knew or assumed. Notably, two Meta 30(b)(6) witnesses testified that Meta has used "torrenting" to acquire millions of copyrighted works from pirated (i.e. illegal) databases. *See, e.g.*, Ex. B at 87-88; Ex. C at 348–351. In the process of torrenting this massive data, moreover, Meta also "seeded" it to others in the online piracy community. Ex. C at 348–351. The full extent of Meta's torrenting formatively bears on Meta's intentional copying and use of pirated books and awareness that this conduct was legally problematic given Meta's efforts to prevent the public from being able to trace its torrenting activity back to Meta IP addresses and Facebook servers. Meta_Kadrey_00204223 (Bashlykov: "not sure we can use meta's IPs to load through torrents pirate content").

Meta's torrenting-related data is thus directly relevant to Plaintiffs' copyright infringement claim because it reflects some of the copyrighted data that Meta downloaded from the shadow/pirated libraries at issue in this case, and it is also evidence of Meta *distributing* this copyrighted data without consent from the actual copyright holders, which is an independent infringing act. To fully assess the scope of Meta's torrenting, Plaintiffs asked Meta to produce its BitTorrent client, application logs, and peer lists—data that reflects how much Meta downloaded and from where, and how much Meta seeded (i.e., reuploaded) to the internet. This data is unquestionably responsive to Plaintiffs' RFP 85 ("All Documents and Communications related to the decision to use Torrent Systems to acquire data for training Llama Models") and RFP 119 ("All Documents and Communications, including source code, relating to the processing of copyrighted material used in training Llama Models, including storage and deletion of copyrighted material."). While Meta has produced *other* responsive torrenting-related documents, Meta refuses to produce the actual data about what was torrented. That is fundamental discovery and the Court should order its production.

B. Supervised Fine-Tuning Data

Mr. Bashlykov also testified that Meta stores its "supervised fine-tuning data" for its Llama models on a specific hard drive cluster referred to as EAG-WSF. Ex. B at 144-146. Plaintiffs have observed what appeared to be gaps in Meta's mitigation data productions, and Plaintiffs now know from Mr. Bashlykov that this set of fine-tuning data exists within a discrete data location that has not been produced. Plaintiffs' RFPs 118 and 119 (already subjects of a pending discovery letter, Dkt. No. 308) also cover this specific fine-tuning dataset. For the same reasons Plaintiffs explained in that letter, Meta's supervised fine-tuning data is relevant. That data regulates Llama by (1) training the model to identify when copyrighted material has been emitted and (2) preparing alternative answers when copyrighted emissions occur. The supervised fine-tuning data consists

of copyrighted works themselves: in short, the model is fine-tuned to say, "Don't emit this." Thus, not only does Meta's supervised fine-tuning data itself consist of Plaintiffs' and putative class members' works, but whether Llama models frequently regurgitate copyrighted material unless fine-tuned also bears on Meta's fair use argument that Llama models' outputs are "transformative."

There is no debate that the supervised fine-tuning data exists. Mr. Bashlykov testified to where it's located. And it is clearly responsive. The Court should therefore compel its production.

C. Llama 4 and 5 Training Datasets

This Court already held that Llamas 4 and 5 are not analytically different from Llamas 1-3. *See* Dkt. No. 279 at 4 ("Llama 4 is relevant to this case, notwithstanding that it is still under development."); Dkt. No. 315 at 7 ("Llama 5 is relevant"). In its order addressing the parties' dispute over the definition of "Shadow Datasets," the Court also stated that "since Meta has disclosed all the datasets that were used from Llama 1, 2, and 3, Plaintiffs can tell Meta which of those are Shadow Datasets." Dkt. No. 315 at 8. Meta's representation that it has fully produced the training datasets for Llamas 1-3 resolves *that* issue. But Meta has not done the same for Llamas 4 or 5, and indeed refuses to do so. Instead, Meta has only produced "documents sufficient to show various other datasets that have been or are being considered for use with future models." *Id.* at 7.

The problem is that Meta's witnesses have testified that additional "Shadow Datasets" were used as Llama 4 training data, and those datasets are not reflected in Llamas 1-3. For example, the evidence makes clear that over the last several months, Meta has increased its reliance on Shadow Datasets, including the notorious "Z-Lib" Shadow Dataset, which had numerous domains seized by the FBI in recent years.[1] Moreover, Meta has begun to source new copyrighted works for its LibGen dataset through the website Anna's Archive. Yet, Meta has not produced any of this data, and Plaintiffs still do not possess the full Llama 4 (or Llama 5) training datasets despite their relevance.

This missing training data plainly exists. When Plaintiffs emailed Meta to confirm whether it had produced the entire training datasets for Llamas 4 and 5, Meta responded with a bizarre objection that the request does not correspond to any RFPs. Meta's response clearly indicates that Meta possesses more Llama 4-5 training data; if Meta did not, it would have said that all responsive data was produced. Further, Meta's contention is untrue: the Llama 4 training dataset is responsive to multiple RFPs. RFP 81 asks for "All Documents and Communications related to the decision to use Shadow Datasets for training Llama Models," which would encompasses the actual training data pulled from the Shadow Datasets (certainly the actual datasets are "related to" the decision to use those datasets). Dkt. 294-1 at 8-9. RFPs 6-12 also cover all documents and communications with enumerated Shadow Datasets concerning training data. And finally, the fact that Meta purportedly produced documents "sufficient to show" the datasets under consideration for use with future models, Dkt. No. 315 at 7, is a concession of relevance—Meta would not have produced "sufficient to show" documents that fall outside its interpretation of Plaintiffs' discovery requests.

Accordingly, Plaintiffs request an order compelling production of Meta's complete training datasets used for Llama 4 (and Llama 5 if any datasets exists), and at minimum, the training data subsets derived from the sources listed in Plaintiffs' definition of "Shadow Datasets": Books3, Z-

[1] https://techhq.com/2023/11/how-is-z-library-down-again-alternatives-ebooks/

Library (aka B-ok), Library Genesis, Bibliotik, Anna s Archive, and The Pile. Dkt. No. 294-1 at 4.

META'S STATEMENT

Documents Related to Alleged "Torrenting" In this motion, Plaintiffs distort deposition testimony to once again demand documents they did not ask for in discovery and that are not relevant to the lone remaining copyright infringement claim—namely, Meta's alleged "BitTorrent client" software, application logs from the alleged use of torrenting, or peer lists created during the alleged use of torrenting. Plaintiffs did not ask for this information in their document requests. And Plaintiffs readily admit that Meta already has produced documents regarding discussions of alleged torrenting within Meta, including documents regarding any alleged decisions to use torrents to acquire data for training the Llama models. There is nothing to compel here.

Plaintiffs cite two RFPs in their brief, Nos. 85 and 119, but neither calls for any of the documents Plaintiffs now demand. RFP 85 is specifically directed to decision-making about the alleged use of torrenting to acquire data for training Llama models: "All Documents and Communications related to the *decision to use* Torrent Systems to acquire data for training Llama Models." Ex. D (RFP 85). Contrary to Plaintiffs' assertion, RFP 85 does not mention and cannot plausibly be read to request "actual data about what was torrented," much less any alleged "BitTorrent client," application logs, or peer lists.

The Court has already ruled that RFP 119 does not cover requests for documents about torrenting. ECF 351 at 2-3 ("The Court does not see how torrenting is responsive to this RFP, which is about the processing of data, not its acquisition"). Consistent with the Court's prior ruling, as much as Plaintiffs would like to use this RFP as a catchall for anything they might want in the moment, it makes no reference to torrenting or other means of *acquiring* data for training the Llama models. Rather, it seeks documents relating to the *processing* of copyrighted material used in training, including the storage and deletion of such material.

If Plaintiffs had intended to request production of training data Meta acquired through torrents, there was a simple way to do that. Plaintiffs made no such request. Indeed, the application logs and peer lists sought by Plaintiffs are the types of materials that the parties agreed in the ESI order are not proportional to the needs of the case. ECF 101 at 7. Application logs and peer lists would at least qualify as "[s]erver, system, or network logs" and "[o]n-line data such as temporary internet files, history, cache, cookies, and the like," which the ESI order exempts from production in this case. ECF 101 at 7 (¶ 8.C.iii & vi).

Supervised Fine-Tuning Data. Plaintiffs admit that they have already asked for this material in their pending motion on RFPs 118-119, which was heard at the December 19, 2024 hearing, with the Court asking for supplemental briefing with a narrower "ask" from Plaintiffs. Even if the Court were willing to entertain yet another motion on these same RFPs (it should not), neither supports Plaintiffs' request for production of "fine-tuning data."

RFP 118 seeks documents "relating to any efforts, attempts, or measures implemented by Meta to prevent Llama Models from emitting or outputting copyrighted material." This RFP is directed to documents describing or showing "efforts, attempts, and measures," which Meta has produced. It does ***not*** ask for training data of any kind.

Likewise, the Court has already ruled that RFP 119 does not cover datasets or copies of copyrighted works. ECF 351 at 2-3 (noting that RFP 119 "did not ask for" "datasets that include

Plaintiffs' copyrighted works"). As the Court knows, RFP 119 seeks "All Documents and Communications, including source code, relating to the ***processing*** of copyrighted material used in training Llama Models, including storage and deletion of copyrighted material." As the Court has previously recognized, Plaintiffs cannot morph a request for documents describing the processing, storage, and deletion of copyrighted material into a request for post-training data they never asked for. Additionally, as with the prior motion, Plaintiffs again baldly assert that fine-tuning data "bears on" the issue of transformativeness for fair use, but again provides no explanation why or how it is connected to the issues in dispute.

Finally, Plaintiffs' citations to the Bashlykov deposition do not support their argument that any of this material is relevant here. The muddled questioning does not evidence any specific tie between the fine-tuning data Mr. Bashlykov is talking about and any copyrighted works, let alone the Plaintiffs' works, nor any connection to any alleged copyright mitigation efforts. *See* Ex. B at 144-46. Plaintiffs' motion on supervised fine-tuning data should be denied.

Llama 4 and 5 Training Datasets. Plaintiffs did not ask for any training datasets in discovery other than the datasets for Llamas 1-3. Indeed, Plaintiffs knew how to specifically ask for training datasets; their first three RFPs served on December 27, 2023 specifically asked for "The Training Data for Llama [1/2/3]."[2] Ex. E (excerpt of RFPs 1-3). And there is no dispute that Meta produced that data in the case, subject to compromises between the Parties, due to the burden of collecting these datasets, to limit the data for Llama 3 to datasets relating to books, similar to Plaintiffs' requested "minimum relief" here. Plaintiffs' RFPs specifically asked for training data for Llama 1, 2, and 3 *only*. They never asked for the training data for Llama 4 or Llama 5.

In the absence of specific RFPs directed to Llama 4 or 5 training data, Plaintiffs cite RFPs 6-12 and 81. To start, RFPs 6-12 (like RFPs 1-3, which Plaintiffs notably do not reference) are Existing Written Discovery served nearly a year ago, and any motion practice directed to those RFPs was required to be filed no later than Nov. 8, 2024, as extended by Judge Chhabria following Plaintiffs failure to comply with the Court's earlier deadline. ECF 253. Plaintiffs' reliance on these RFPs as a basis for its current motion to compel disregards Judge Chhabria's orders. Moreover, these Requests seek communications with various alleged "organizations," not the datasets themselves.

Plaintiffs also cite to RFP 81, but that RFP does not support the production of "all" Llama 4 and Llama 5 training data. Like RFP 85 discussed above, RFP 81 is specifically directed to materials about *decision-making* surrounding the alleged use of "Shadow Datasets," not the datasets themselves: "All Documents and Communications *related to the decision to use* Shadow Datasets for training Llama Models" And even if this RFP could be distorted to cover actual training data (which is should not), Plaintiffs do not even attempt to explain how RFP 81 might require production of "Meta's complete training datasets used for Llama 4 (and Llama 5 if any datasets exists)", which Plaintiffs concede are not limited to so-called "Shadow Datasets."

Finally, producing all training data (let alone multiple copies) for Llama 4 would be exceptionally burdensome and not proportional to the needs of the case. Meta has investigated the burden involved and just the datasets in Plaintiffs' requested "minimum" relief of identified "Shadow Datasets," when exported, would result in ~48 million rows of data across 9 data tables, comprising ~5 terabytes of data (size estimated based on a sampling procedure). Meta estimates that, assuming

[2] Unlike other requests asking about "Llama Models" generally, where the Court found Llama 4 and 5 to be relevant, these Requests seek documents re these specific Llama models only.

there are no unanticipated issues, for just three of the data tables it would take at least five weeks to complete the export process alone. The export process for those three tables is anticipated to take at least five weeks because Meta would need to engage its data scientists to write and test custom software code to break up the data into smaller portions. This is an iterative process that involves intermediate testing and quality assurance steps until all the data has been exported. It will also take several weeks of machine processing time to export the data into a format that can be produced for review. Export of the other six data tables in the "minimum" relief requested by Plaintiffs is estimated by Meta to take at least ten weeks. Had Plaintiffs' timely and explicitly requested this, as they did for Llama 1-3 (RFPs 1-3), Meta could have produced this data, but this eleventh-hour ask after the close of discovery makes this infeasible. Moreover, all training datasets that may be used for Llama 4 (and to the extent known for Llama 5) would constitute more than a hundred data tables and it is unclear that this amount of data could even be exported, and if so, how long it might take.

Plaintiffs' demands for just the alleged "Shadow Datasets" would be incredibly burdensome exercise that would provide minimal if any relevant information for the issues in this case, as Plaintiffs' works, to the extent they are included in the training data, are a miniscule portion of the overall training data set, and Meta has already admitted that text from each of the Plaintiffs' books was included in the Books3 dataset used to train Meta's Llama models. (*See* amended responses to RFAs 3–6, ECF 352, Ex. A.) Meta has also already produced documentation to Plaintiffs that identifies the datasets that are being used for ongoing Llama 4 training. Llama 5 remains at early stages of planning and it is not yet known what datasets will be used for training at this time.

III. PLAINTIFFS' REPLY

With respect to another dispute, this Court did not order production of torrent-related discovery. Dkt. 351 at 2–3. Yet, as is more patently relevant here, RFP 119 concerns "processing." Indeed, torrenting is not acquisition but *processing*. A party must *process* large torrented files into "chunks" which are then *reprocessed* into their original files. Thus, torrenting ***is*** also an act of data processing.[3] Understanding how Meta processed this data will speak to what was copied and how. It will also demonstrate how, to whom, and what exactly was distributed. Similarly, "application logs" and "peer lists" reflect a set of the servers "storing" data, responding to the "storage" element.

With respect to Meta's decision to torrent pirated data (RFP 85), *how* Meta torrented squarely pertains to that *decision*, because some methods would have more readily facilitated or concealed Meta's infringement. The more we know about that, the more we can understand why Meta *decided* to use which types of torrenting (relevant, at minimum, to willfulness and bad faith). Here, Meta has produced fewer than 30 documents discussing torrenting. Yet, Meta deponents have testified that torrenting was and remains a key method used to copy millions of pirated works.

Meta should also produce the Shadow Library training datasets used for training Llamas 4 and 5—much of which it copied well before the launch of Llama 3. "Communications" with these Libraries (including direct downloads which are, by definition, communications) encompasses the training data Meta copied from them. Meta cannot claim burden simply because it stole too much data.

[3] Ask Meta.AI, "Does bit torrent involve data processing," and it replies, "Yes, BitTorrent involves data processing in various ways." *See ME2 Prods., Inc. v. Bayu*, 2017 WL 5165487, at *1-2 (D. Nev. Nov. 7, 2017) (re: processing); *New Sensations, Inc. v. Does 1-426*, 2012 WL 4675281, at *2 (N.D. Cal. Oct. 1, 2012) (Corley, J.) (same); *see also* https://medium.com/@kyodo-tech/the-bittorrent-dht-and-decentralized-content-sharing-bb91befdb294.

By: /s/ *Bobby Ghajar*

Bobby A. Ghajar
Colette A. Ghazarian
COOLEY LLP
1333 2nd Street, Suite 400
Santa Monica, CA 90401
Telephone: (310) 883-6400
Facsimile: (310) 883-6500
Email: bghajar@cooley.com
cghazarian@cooley.com

Mark R. Weinstein
Elizabeth L. Stameshkin
COOLEY LLP
3175 Hanover Street
Palo Alto, CA 94304
Telephone: (650) 843-5000
Facsimile: (650) 849-7400
Email: mweinstein@cooley.com
lstameshkin@cooley.com

Kathleen R. Hartnett
Judd D. Lauter
COOLEY LLP
3 Embarcadero Center, 20th Floor
San Francisco, CA 94111
Telephone: (415) 693-2071
Facsimile: (415) 693-2222
Email: khartnett@cooley.com
jlauter@cooley.com

Phillip Morton
COOLEY LLP
1299 Pennsylvania Avenue, NW, Suite 700
Washington, DC 20004
Telephone: (202) 842-7800
Facsimile: (202) 842-7899
Email: pmorton@cooley.com

Angela L. Dunning
CLEARY GOTTLIEB STEEN & HAMILTON LLP
1841 Page Mill Road, Suite 250
Palo Alto, CA 94304
Telephone: (650) 815-4121

By: /s/ *Maxwell V. Pritt*

BOIES SCHILLER FLEXNER LLP
David Boies (*pro hac vice*)
333 Main Street
Armonk, NY 10504
(914) 749-8200
dboies@bsfllp.com

Maxwell V. Pritt (SBN 253155)
Joshua M. Stein (SBN 298856)
44 Montgomery Street, 41st Floor
San Francisco, CA 94104
(415) 293-6800
mpritt@bsfllp.com
jstein@bsfllp.com

Jesse Panuccio (*pro hac vice*)
1401 New York Ave, NW
Washington, DC 20005
(202) 237-2727
jpanuccio@bsfllp.com

Joshua I. Schiller (SBN 330653)
David L. Simons (*pro hac vice*)
55 Hudson Yards, 20th Floor
New York, NY 10001
(914) 749-8200
dsimons@bsfllp.com
jischiller@bsfllp.com

Interim Lead Counsel for Plaintiffs

Facsimile: (650) 849-7400
Email: adunning@cgsh.com

Attorneys for Defendant Meta Platforms, Inc.

ATTESTATION PURSUANT TO CIVIL LOCAL RULE 5-1(h)

I hereby attest that I obtained concurrence in the filing of this document from each of the other signatories. I declare under penalty of perjury that the foregoing is true and correct.

Dated: December 20, 2024

BOIES SCHILLER FLEXNER LLP

/s/ Maxwell V. Pritt
Maxwell V. Pritt
Reed Forbush
Jay Schuffenhauer

Attorneys for Plaintiffs